LEAKY TILING ACTIVATIONS: A SIMPLE APPROACH TO LEARNING SPARSE REPRESENTATIONS ONLINE

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work has shown that sparse representations—where only a small percentage of units are active—can significantly reduce interference. Those works, however, relied on relatively complex regularization or meta-learning approaches, that have only been used offline in a pre-training phase. We design an activation function that naturally produces sparse representations, and so is more amenable to online training. The idea relies on the simple approach of binning, but overcomes the two key limitations of binning: zero gradients for the flat regions almost everywhere, and lost precision—reduced discrimination—due to coarse aggregation. We introduce a Leaky Tiling Activation (LTA) that provides non-negligible gradients and produces overlap between bins that improves discrimination. We first show that LTA is robust under covariate shift in a synthetic online supervised problem, where we can vary the level of correlation and drift. Then we move to deep reinforcement learning setting and investigate both value-based and policy gradient algorithms that use neural networks with LTAs, in classic discrete control and Mujoco continuous control environments. We show that algorithms equipped with LTAs are able to learn a stable policy faster without needing target networks on most domains.

1 INTRODUCTION

Representation learning in online reinforcement learning (RL) systems can strongly impact learning efficiency, both positively due to generalization but also negatively due to interference (Liang et al., 2016; Heravi, 2019; Le et al., 2017; Liu et al., 2019; Chandak et al., 2019; Caselles-Dupré et al., 2018; Madjiheurem & Toni, 2019). Neural networks particularly suffer from interference—where updates for some inputs degrade accuracy for others—when training on temporally correlated data (McCloskey & Cohen, 1989; French, 1999; Kemker et al., 2018). Data is naturally temporally correlated in online RL. The effect might even be worse in RL, because the effect of interference is likely magnified by bootstrapping. Target networks—a slowly-updating network used to compute the bootstrap target (Mnih et al., 2015)—can mitigate some of this interference and stabilize training (van Hasselt et al., 2018; Mnih et al., 2015; Yang et al., 2019), but can further slow learning by using outdated estimates of value in the targets (Ghiassian et al., 2020).

Recent work (Liu et al., 2019; Ghiassian et al., 2020; Javed & White, 2019; Rafati & Noelle, 2019; Hernandez-Garcia & Sutton, 2019), as well as older work (McCloskey & Cohen, 1989; French, 1991), have shown that sparse representation can reduce interference in updates. A sparse representation is one where only a small number of features are active, for each input. Each update only impacts a small number of weights and so is less likely to interfere with many state values. Further, when constrained to learn sparse features, the feature vectors are more likely to be orthogonal (Cover, 1965), which further mitigates interference. The learned features can still be highly expressive, and even more interpretable, as only a small number of attributes are learned to describe an input.

Learning sparse representations online, however, remains relatively open. Some previous work has relied on representations pre-trained before learning, either with regularizers that encourage sparsity (Liu et al., 2019) or with meta-learning (Javed & White, 2019). Other work has trained the sparse-representation neural network online, by using sparsity regularizers online with replay buffers (Hernandez-Garcia & Sutton, 2019) or using a winner-take-all strategy (Rafati & Noelle, 2019). Hernandez-Garcia & Sutton (2019) found that many of these sparsity regularizers were ineffective for obtaining sparse representations without high levels of dead neurons, though the regularizers did still

often improve learning. The Winner-Take-All approach is non-differentiable, and there are mixed results on it's efficacy, some positive (Rafati & Noelle, 2019) and some negative (Liu et al., 2019). Providing a simpler approach to obtain sparse representations, that are easy to train online, would make it easier for RL researchers to adopt sparse representations and further explore their utility.

In this work, we introduce a Leaky Tiling Activation function that makes it easy to control the sparsity level, and is simple to train. The idea relies on re-expressing a binning operation—where inputs are aggregated into intervals—which we call a Tiling Activation. Using this functional form, we introduce a smoothing parameter η that ensures gradient information can flow backwards through the activation, while still maintaining the sparsity obtained by binning. We prove that the LTA guarantees sparsity by construction. We first investigate the properties in an online supervised learning problem, where we can carefully control the level of correlation. We then empirically show that RL algorithms using LTA learn more quickly and stably compared to both those with standard ReLU activations and several online sparse representation learning approaches.

2 Problem Formulation

LTA is a generic activation that can be applied in a variety of settings. The motivation for the activation, however, is for the online setting, for temporally correlated data streams. We therefore explicitly formalize the two problems we are motivated by: the online supervised learning problem and the reinforcement learning (RL) problem.

In online supervised learning, the agent observes a temporally correlated stream of data, generated by a stochastic process $\{(X_t, Y_t)\}_{t\in\mathbb{N}}$. The observations X_t depend on the past X_{t-1}, X_{t-2}. The target Y_t depends only on X_t, with a stationary underlying mean function $f(x) = \mathbb{E}[Y_t|X_t = x]$. On each time step, the agent observes X_t, makes a prediction $f_\theta(X_t)$ with its parameterized function f_θ, receives target Y_t and incurs a squared prediction error.[1] The goal of the agent is to minimize prediction error, where f is the ideal predictor. Learning is complicated by the correlation in the data, unlike standard supervised learning where data is independent and identically distributed (iid).

In the RL setting, we formalize the interaction using Markov decision processes (MDPs). An MDP consists of $(\mathcal{S}, \mathcal{A}, \mathbb{P}, R, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathbb{P}$ is the transition probability kernel, R is the reward function, and $\gamma \in [0, 1]$ is the discount factor. At each time step $t = 1, 2, \ldots$, the agent observes a state $s_t \in \mathcal{S}$ and takes an action $a_t \in \mathcal{A}$. Then the environment transits to the next state according to the transition probability distribution, i.e., $s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)$, and the agent receives a scalar reward $r_{t+1} \in \mathbb{R}$ according to the reward function $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$. A policy is a mapping from a state to an action (distribution) $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$. For a given state-action pair (s, a), the action-value function under policy π is defined as $Q_\pi(s, a) = \mathbb{E}[G_t|S_t = s, A_t = a; A_{t+1:\infty} \sim \pi]$ where $G_t \stackrel{\text{def}}{=} \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1})$ is the return of a sequence of transitions $s_0, a_0, s_1, a_1, \ldots$ by following the policy π.

The goal of an agent is to find an optimal policy π^*: the policy that obtains maximal expected return from each state. The policy is either directly learned, as in policy gradient methods (Sutton et al., 1999; Sutton & Barto, 2018), or the action-values are learned and the policy inferred by acting greedily with respect to the action-values, as in Q-learning (Watkins & Dayan, 1992). In either setting, we often parameterize these functions as a Neural Network (NN), $\pi_\theta : \mathcal{S} \to \mathcal{A}$, which inputs a given state (or given observation), and outputs an action: $A_t = \pi_\theta(S_t)$. An action-value $Q_\theta : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ can also be parameterized using a NN, typically by inputting states into the network and outputting multiple values, one for each possible action.

3 Binning with Non-negligible Gradients

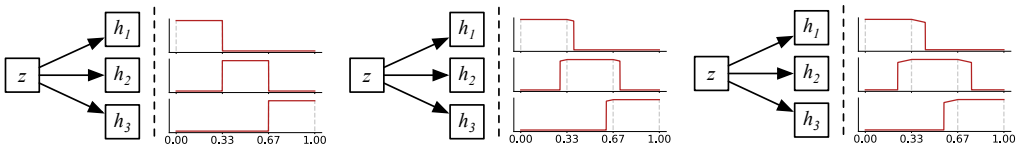
In this section, we develop the *Leaky Tiling Activation* (LTA), as a new modular component for neural networks that provides sparse layers. We first introduce a new way to compute the binning of an input, using indicator functions. This activation provides guaranteed sparsity—only one unit

[1] More generally, we can use other prediction errors, including classification errors. Correspondingly, the ideal prediction f would be different, corresponding to the minimum of this prediction error. For simplicity, we restrict our attention to squared errors, where the ideal prediction is the mean.

is active—but has a gradient of zero almost everywhere. Then, building on this form, we provide a smoothed version, where activation is leaked into nearby bins resulting in non-negligible gradients that make it compatible with back-propagation algorithms. We then prove that the leaky version is still guaranteed to have sparse activation, with the level controllable by the parameter η, which controls how much activation leaks to nearby units.

3.1 Tiling Activation

The tiling activation inputs a scalar z and outputs a binned vector. This vector is one-hot, with a 1 in the bin corresponding to the value of z, and zeros elsewhere. Note that a standard activation typically maps a scalar to a scalar. However, the tiling activation maps a scalar to a vector, as depicted in Figure 1(a). This resembles tile coding, which inspires the name Tiling Activation; to see this connection, we include a brief review of tile coding in the Appendix A.1. In this section, we show how to write the tiling activation compactly, using element-wise max and indicator functions.



(a) TA, $k = 3$ (b) LTA, $k = 3, \eta = 0.05$ (c) LTA, $k = 3, \eta = 0.1$

Figure 1: a) The regular TA mapping $\mathbb{R} \to \mathbb{R}^k$, with each output element h_i corresponds to a different bin. b) The LTA with $\eta > 0$, permitting both overlap in activation, and nonzero gradient between the vertical red and gray lines. c) Larger values for η extends the sloped lines further from either side of each plateau, increasing the region that has non-negligible gradients.

Assume we are given a range $[l, u]$ for constants $l, u \in \mathbb{R}$, where we expect the input $z \in [l, u]$. The goal is to convert the input, to a one-hot encoding, with evenly spaced bins of size $\delta \in \mathbb{R}^+$. Without loss of generality, we assume that $u - l$ is evenly divisible by δ; if it is not, the range $[l, u]$ could be slightly expanded, evenly on each side, to ensure divisibility. Define the k-dimensional tiling vector

$$\mathbf{c} \stackrel{\text{def}}{=} (l, l + \delta, l + 2\delta, ..., u - 2\delta, u - \delta). \tag{1}$$

where $k = (u - l)/\delta$. The **tiling activation** is defined as

$$\boldsymbol{\phi}(z) \stackrel{\text{def}}{=} 1 - I_+(\max(\mathbf{c} - z, 0) + \max(z - \delta - \mathbf{c}, 0)) \tag{2}$$

where $I_+(\cdot)$ is an indicator function, which returns 1 if the input is positive, and zero otherwise. The indicator function for vectors is applied element-wise. In the following theorem, we prove that $\boldsymbol{\phi}$ returns a binned encoding: if $\mathbf{c}_i < z < \mathbf{c}_{i+1}$, then $\boldsymbol{\phi}(z)$ returns $\mathbf{e}_i$ the one-hot (standard basis) vector with a 1 in the i-th entry and zero elsewhere. For values of z that fall on the boundary, $z = \mathbf{c}_i$, the encoding returns a vector with ones in both entries $\mathbf{c}_{i-1}$ and $\mathbf{c}_i$. We provide the following example for intuition for the tiling activation.

Example. Assume $[l, u] = [0, 1]$ and set the tile width to $\delta = 0.25$. Then the tiling vector $\mathbf{c}$ has four tiles ($k = 4$): $\mathbf{c} = (0, 0.25, 0.5, 0.75)$. If we apply the tiling activation to $z = 0.3$, because $0.25 < 0.3 < 0.5$, the output should be $(0, 1, 0, 0)$. To see $\boldsymbol{\phi}(z)$ does in fact return this vector, we compute each max term

$$\max(\mathbf{c} - z, 0) = (0, 0, 0.2, 0.45) \quad \text{and} \quad \max(z - \delta - \mathbf{c}, 0) = \max(0.05 - \mathbf{c}, 0) = (0.05, 0, 0, 0).$$

The addition of these two is $(0.05, 0, 0.2, 0.45)$ and so $1 - I_+(0.05, 0, 0.2, 0.45) = 1 - (1, 0, 1, 1) = (0, 1, 0, 0)$. The first max extracts which components in $\mathbf{c}$ are strictly greater than z, and the second max those which are strictly less than. The addition gives the bins that are strictly greater and strictly less than the bin for z, leaving only the entry corresponding to that bin as 0, with all others positive.

Going beyond this intuition, we can formally characterize the three possible situations for the LTA: 1) it gives exactly one nonzero entry when the input is greater than any element in $\mathbf{c}$; 2) it gives two nonzero entries when the input value is equal to one of the element in $\mathbf{c}$ except the first one; and 3) it gives exactly one nonzero entry when the input is equal to the first element in $\mathbf{c}$. See Appendix A.2.1 for a simple proof. We assume throughout that δ evenly divides the range $[l, u]$; we state this explicitly in Assumption 1 to highlight this condition. This is not restrictive and can be easily satisfied simply by expanding the range slightly, beyond the actual range of inputs.

3.2 LEAKY TILING ACTIVATION (LTA)

The Tiling Activation provides a way to obtain sparse, binary encodings for features learned within a neural network. As motivated upfront, such encodings can enable faster learning and reduce interference. Unfortunately, the tiling activation has a zero derivative almost everywhere. The value switches from 0 to 1, when a threshold is passed (as shown in Figure 1(a)). In this section, we provide a leaky tiling activation, that has non-zero derivatives and so is amenable to use with backpropagation.

To design the LTA, we define the leaky indicator function

$$I_{\eta,+}(x) \stackrel{\text{def}}{=} I_+(\eta - x)x + I_+(x - \eta) \tag{3}$$

where η is a small constant for controlling the sparsity. The first term $I_+(\eta - x)$ is 1 if $x \leq \eta$, and 0 otherwise. The second term $I_+(x - \eta)$ is 1 if $x > \eta$, and 0 otherwise. If $x \leq \eta$, then $I_{\eta,+}(x) = x$, and else $I_{\eta,+}(x) = 1$. The original indicator function I_+ can be acquired by setting $\eta = 0$. When $\eta > 0$, the derivative is non-zero for $x < \eta$, and zero otherwise. This means the derivative can be propagated backwards through this indicator, for this range of inputs, similar to the ReLU activation. Using this leaky indicator function, we define the following **Leaky Tiling Activation** (LTA)

$$\boldsymbol{\phi}_\eta(z) \stackrel{\text{def}}{=} 1 - I_{\eta,+}(\max(\mathbf{c} - z, 0) + \max(z - \delta - \mathbf{c}, 0)) \tag{4}$$

where again $I_{\eta,+}$ is applied elementwise.

We depict this activation, with two different η, in Figures 1(b) and 1(c). For the smaller η, the LTA extends the activation to the neighbouring bins. The activation in these neighbouring bins is sloped, resulting in non-zero derivatives. For this smaller η, however, there are still regions where the derivative is zero (e.g., $z = 0.3$ in Figure 1(b)). The regions where derivatives are non-zero can be expanded by increasing η as shown in Figure 1(c). Hence we can adjust η to make trade-off between the sparsity level and how much neural network can change as we demonstrate in Section A.5. Figure 2 shows a neural network with LTA applied to the second hidden layer and its output y is linear in the sparse representation. LTA itself does not introduce any new training parameters, just like other activation functions. For input $\mathbf{x}$, after computing first layer $\mathbf{h}_1 = \mathbf{x}\mathbf{W}_1$, we apply $\boldsymbol{\phi}_\eta(\mathbf{z})$ to $\mathbf{h}_1\mathbf{W}_2 \in \mathbb{R}^d$ to get the layer $\mathbf{h}_2$ of size kd. This layer consists of stacking the k-dimensional sparse encodings, for each element in $\mathbf{h}_1\mathbf{W}_2$.

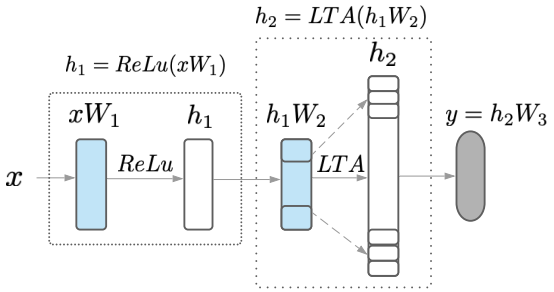


Figure 2: A visualization of an LTA layer

3.3 GUARANTEED SPARSITY FROM THE LTA

We now show that the LTA maintains one of the key properties of the tiling activation: sparsity. The distinction with many existing approach is that our sparsity is guaranteed by design and hence is not a probabilistic guarantee. We first characterize the vectors produced by the LTA, in Proposition 2 with proof in Appendix A.2.2. Then, we provide an upper bound on the proportion of nonzero entries in the generated vector in Theorem 1, with in Appendix A.2.3.

Assumption 1. *$\delta < u - l$, where $k\delta = u - l$ for $k \in \mathbb{N}$.*

Theorem 1 (Sparsity guarantee for LTA.)**.** *For any $z \in [l, u], \eta > 0$, $\boldsymbol{\phi}_\eta(z)$ outputs a vector whose number of nonzero entries $\|\boldsymbol{\phi}_\eta(z)\|_0$ satisfies:*

$$\|\boldsymbol{\phi}_\eta(z)\|_0 \leq 2\left\lfloor \frac{\eta}{\delta} \right\rfloor + 3$$

Corollary 1. *Let $\rho \in [0, 1)$ be the desired sparsity level: the maximum proportion of nonzero entries of $\boldsymbol{\phi}_\eta(z), \forall z \in [l, u]$. Assume $\rho k \geq 3$, i.e., some inputs have three active indices or more (even with $\eta = 0$, this minimal active number is 2). Then η should be chosen such that*

$$\left\lfloor \frac{\eta}{\delta} \right\rfloor \leq \frac{k\rho - 3}{2} \quad \textit{or equivalently} \quad \eta \leq \frac{\delta}{2}\left(\lfloor k\rho \rfloor - 1\right) \tag{5}$$

As an example, for $k = 100$, $\delta = 0.05$ and a desired sparsity of at least 10% ($\rho = 0.1$), we can use $\eta = \frac{0.05}{2}(\lfloor 100 \times 0.1 \rfloor - 1) = 0.225$. Note that the bound in Theorem 1 is loose in practice as the bound is for any input z. In Appendix A.2.3 and A.5, we theoretically and empirically show that the actual sparsity is usually lower than the upper bound, and quite consistent across inputs.

4 EXPERIMENTS IN SUPERVISED LEARNING UNDER COVARIATE SHIFT

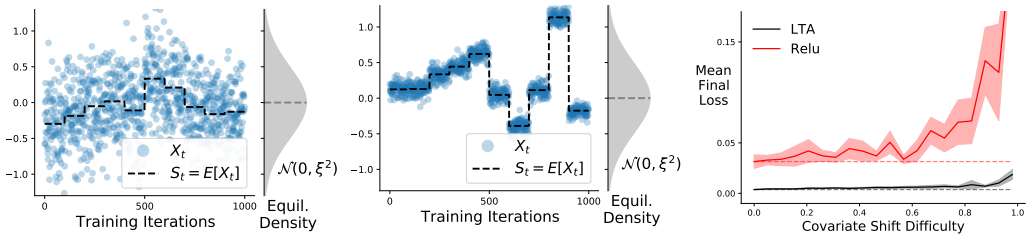
A key property of LTA is that it provides immediate, deterministic sparsity guarantee. We hypothesize that such property is suitable to handle highly nonstationary data, where there is a strong need for interference reduction. As a result, in this section, we focus on testing the hypothesis that LTA provides representations that are more robust to learning online on correlated data, Specifically, we hypothesize that convergence speed and stability for ReLU networks suffer under strongly correlated training data, whereas comparable LTA networks are nearly unaffected. We create a synthetic problem with a relative simply target function, to focus the investigation on the role of a drifting distribution on inputs, which results both in covariate shift and creates temporal correlation during training. We also report results on two benchmark image classification tasks in the Appendix A.6.

The Piecewise Random-Walk Problem has Gaussian $X_t \sim \mathcal{N}(S_t, \beta^2)$ with fixed variance β^2 and a mean S_t that drifts every T steps, with the target $y_t = \sin(2\pi x_t^2)$ where x_t is a realization of X_t. More precisely, the mean S_t stays fixed for T timesteps, then takes a step according to a first order autoregressive random walk: $S_t = (1-c)S_t + Z_t$ where $c \in (0, 1]$ and $Z_t \sim \mathcal{N}(0, \sigma^2)$ for fixed variance σ^2. If $c = 0$, then this process is a standard random walk; otherwise, with $c < 1$, it helps keep S_t in a bounded range with high probability.

This process is designed to modulate the level of correlation—which we call *correlation difficulty*—without changing the equilibrium distribution over X_t. As the correlation difficulty d varies from 0 to 1, the training data varies from low to high correlation; $d = 0$ recovers iid sampling. All $d \in [0, 1)$ share the same equilibrium distribution in X_t. X_t is ergodic and has Gaussian equilibrium distribution $\mathcal{N}(0, \xi^2)$, with variance ξ^2 dependent upon β^2, σ^2 and c. In particular, the visitation distribution X_t for any training run will converge to the equilibrium distribution. This ensures that when we measure online prediction error, the visitation distribution over X_t is very similar across settings. We depict sample trajectories with low d and high d in Figure 3(a). For a rigorous construction of X_t and S_t, and justification for this equilibrium distribution and implementation details, see Appendix A.7.

We measure the mean squared error over the equilibrium distribution in X_t, for neural networks using LTA and ReLU activations across a range of correlation difficulty values. In Figure 17, we can see that LTA outperforms ReLU in two ways. First, LTA converges to a lower loss with less variance across all correlation difficulties. Second, LTA only marginally suffers under high difficulties $d > 0.9$, whereas the performance of ReLU begins to deteriorate for relatively mild $d > 0.5$.

One potentially surprising point is that the LTA reaches a lower error, even on iid data ($d = 0$). We chose the size of the ReLU neural network based on best performance on iid data, meaning it does not have precisely the same number of parameters as LTA. As a sanity check, we additionally include results for ReLU with exactly the same number of parameters as the LTA, in the Appendix in Figure 19. These results are very similar to the ReLU included in Figure 3(c). We hypothesize that the gap between the ReLU and LTA even on iid data arises due to the fact that they are training online with one sample at a time, rather than larger mini-batches.



(a) Mild Correlation (b) High Correlation (c) ReLU and LTA, $d \in [0, 1)$

Figure 3: (a) and (b) contain sample trajectories of X_t which are (a) mildly correlated with $d = 0.41$ and (b) severely correlated with $d = 0.98$. Both share the same equilibrium distribution (in gray). (c) A plot of the prediction error (under the equilibrium distribution), averaged over the final 2.5K iterations, across a range of difficulty settings. All networks are trained for 20k online updates. The lines correspond to the mean of 10 runs with the shaded region corresponding to 95% confidence intervals. The i.i.d setting $d = 0$ is shown as a dotted line for baseline comparison.

5 EXPERIMENTAL RESULTS IN RL

In this section, we empirically study the effect of using LTA in RL. We show overall performance on several benchmark discrete action and continuous action environments. We then compare our method with other simple strategies to obtain sparse representations. Appendix A.4 includes details for reproducing experiments and Appendix A.5 has additional RL experiments.

5.1 ALGORITHMS AND NAMING CONVENTIONS

All the algorithms use a two-layer neural network, with the primary difference being the activation used on the last layer. See Appendix A.5 for results using LTA in all the hidden layers. DQN is used for the discrete action environments, and Deep Deterministic Policy Gradient (DDPG) (Lillicrap et al., 2016) for continuous action environments. On each step, all algorithms sample a mini-batch size of 64 from an experience replay buffer with maximum size 100k. Note that we are able to keep the same LTA setting across *all* experiments: we set $[l, u] = [-20, 20]$; we set $\delta = \eta = 2.0$, $\mathbf{c} = \{-20, -18, -16, ..., 18\}$, and hence $k = 40/2 = 20$.

We first compare to standard DQN agents, with the same architectures except the last layer.
DQN: DQN with tanh or ReLU on the last layer (best performance reported).
DQN-LTA: DQN with LTA on the last layer.
DQN-Large: DQN, but with the last layer of the same size as DQN-LTA. If DQN has a last layer of size d, then DQN-LTA has a last layer of size dk. The number of parameters is actually the same; the LTA activation simply expands the number of features due to binning. We include DQN-Large for a comparison to a DQN agent with the same number of (ReLU or tanh) features in the last layer, used to learn the value function. DQN-Large has k times more parameters in this last layer than DQN or DQN-LTA, and so we expect that over-parameterization to be detrimental.

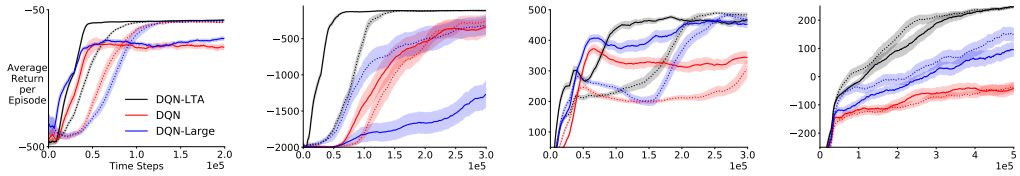
We also compare to several simple strategies to obtain local or sparse features. Radial basis functions (RBFs) have traditionally been used to obtain local features in RL, and recent work has used ℓ_2 and ℓ_1 regularization directly on activations as a simple baseline (Arpit et al., 2016; Liu et al., 2019). All of these strategies have the same sized last layer as the sparse feature dimension of DQN-LTA.
DQN-RBF: DQN using radial basis functions (RBFs) on the last layer, with the centers defined by the same $\mathbf{c}$ as LTA: $\phi(z) = [\exp(-\frac{(z-\mathbf{c}_1)^2}{\sigma}), \ldots, \exp(-\frac{(z-\mathbf{c}_k)^2}{\sigma})]$ where σ is the bandwidth parameter.
DQN-L2/L1: DQN with ℓ_2 or ℓ_1 regularization on the activation functions of the final hidden layer.

To the best of our knowledge, no suitable sparse representation approaches exist for RL. SR-NNs for RL were only developed for offline training (Liu et al., 2019). That work also showed that k-sparse NNs (Makhzani & Frey, 2013) and Winner-Take-All NNs (Makhzani & Frey, 2015) performed significantly worse than ℓ_2 regularization, where ℓ_2 actually performed quite well in most of their experiments. One other possible option is to use Tile Coding NNs (Ghiassian et al., 2020), which first tile code inputs before feeding them into the neural network. This approach focuses on using discretization to break, or overcome, incorrect generalization in the inputs; their goal is not to learn sparse representations. This approach is complementary, in that it could be added to all the agents in this work. Nonetheless, because it is one of the only papers with a lightweight approach to mitigate interference in online RL, we do compare to it in the Appendix A.5.

5.2 OVERALL PERFORMANCE

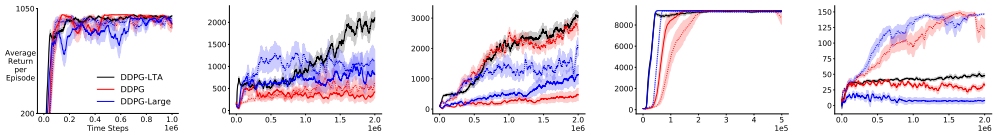
In this section, we demonstrate the overall performance of using LTAs on both discrete and continuous control environments. Our goals are to investigate if we can 1) obtain improved performance with LTA, with fixed parameter choices across different domains; 2) improve stability in learning with LTA; and 3) see if we can remove the need to use target networks with LTA, including determining that learning can in fact be faster without target networks and so that it is beneficial to use LTA without them. All experiments are averaged over 20 runs (20 different random seeds), with offline evaluation performed on the policy every 1000 training/environment time steps.

Discrete control. We compare performance on four discrete-action environments from OpenAI (Brockman et al., 2016): MountainCar, CartPole, Acrobot and LunarLander. We use 64×64 ReLU hidden units on all these domains. Since LTA has $k = 20$, this yields $64 \times 20 = 1280$ dimensional sparse features. Hence, DQN-Large, use two layers ReLU with size 64×1280.



(a) Acrobot (v1) (b) MountainCar (v0) (c) CartPole (v1) (d) LunarLander (v2)

Figure 4: Evaluation learning curves showing episodic return versus environment time steps. The **dotted** line indicates algorithms trained *with* target networks. The results are averaged over 20 runs and the shading indicates standard error. The learning curve is smoothed over a window of size 10 before averaging across runs.



(a) InvertedPendu (b) Hopper (c) Walker2d (d) InvertedDouble (e) Swimmer

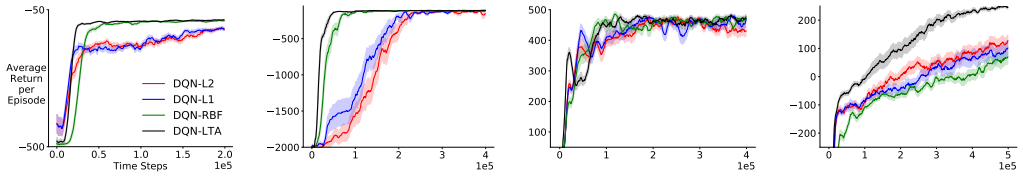
Figure 5: Evaluation learning curves on Mujoco environments, averaged over 20 runs with shading indicating standard error. The **dotted** line indicates algorithms trained *with* target networks. The learning curve is smoothed over a window of size 30 before averaging across runs.

We plot evaluation learning curves, averaged over 20 runs, in Figure 4. The results show the following. 1) With or without using a target network, DQN with LTA can significantly outperform the version without using LTA. 2) LTA has significantly lower variability across runs (smaller standard errors) in most of the figures. 3) DQN-LTA trained *without* a target network outperforms DQN-LTA trained *with* a target network, which indicates a potential gain by removing the target network. 4) Without using LTA, DQN trained without a target network cannot perform well in general, providing further evidence for the utility of sparse feature highlighted in previous works (Liu et al., 2019; Rafati & Noelle, 2019). 5) Simply using a larger neural network does not obtain the same performance improvements, and in some cases significantly degrades performance.

Continuous control. We compare performance on five continuous-action environments from Mujoco (Todorov et al., 2012): Swimmer, Inverted Pendulum, Double Inverted Pendulum, Hopper, and Walker. For these experiments, we use DDPG with exactly the same LTA setting as the above discrete control domain to show its generality. The results are qualitatively similar to the discrete-action environments, as seen in Figure 5, except in one domain (Swimmer). In all other domains, DDPG equipped with LTA, without target networks, achieves comparable and sometimes significantly better performance to DDPG. Swimmer highlights that LTA is not always sufficient on its own to overcome instabilities, and could be complemented by strategies such as using mellowmax (Kim et al., 2019).

5.3 Comparison with Other Sparse Approaches

LTA provides clear benefits, but it is natural to ask if other simple strategies that provide sparse or local features could have provided similar benefits. We compare to DQN-RBF, DQN-L2 and DQN-L1, on the discrete action environments, shown in Figure 6. We find the following. 1) LTA performs consistently well across all environments using a fixed parameter setting; none of the other approaches achieve consistent performance, even though we tuned their parameters per environment. 2) Both the ℓ_1 and ℓ_2 approaches have a high variance across different random seeds. 3) The RBF variant can do better than the ℓ_1 and ℓ_2 approaches but is worse than our algorithm. It is a known issue that RBF is sensitive to the bandwidth parameter choice and we observe similar phenomenon. It is also known that the exponential activation can be problematic in the back-propagation process. We empirically investigate the overlap and instance sparsities of each algorithm in the Appendix A.5.



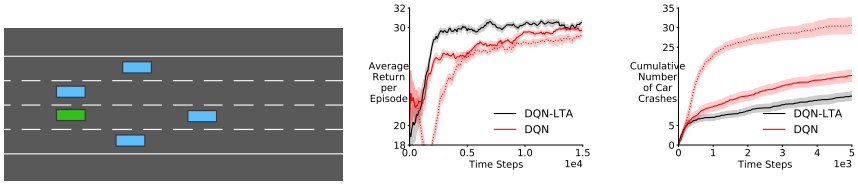
(a) Acrobot (v1) (b) MountainCar (v0) (c) CartPole (v1) (d) LunarLander (v2)

Figure 6: Evaluation learning curves on discrete action domains, averaging over 20 runs with the shade indicating standard error. All algorithms are trained without using target networks. The learning curve is smoothed over a window of size 10 before averaging across runs.

5.4 Testing Stability in a Simulated Autonomous Driving Domain

Our results have shown improved stability with LTA. In this section, we test LTA in an environment focused on stability, namely an autonomous driving task (Leurent, 2018). In the real world, a stable policy is of vital importance to ensure safety in autonomous driving. In this simulated task, the goal is not only to obtain high return, but also keep the number of car crashes as low as possible. Figure 7(a) shows the domain, where the agent—the green car—has to learn to switch lanes, avoid car crashes, and go as fast as possible. The observations are 25-dimensional, with vehicle dynamics that follow the Kinematic Bicycle Model. The action space is discrete.

LTA learns faster, with significantly fewer car crashes incurred during the evaluation time steps as shown in Figure 7(c). Target networks are harmful in this environment, potentially because they slow early learning; the agent to accumulate a significant number of crashes before improving.



(a) Highway (b) Evaluation LC (c) Car Crashes LC

Figure 7: (a) The Highway environment. (b) The evaluation learning curve. (c) The cumulative number of car crashes as a function of driving time steps. Results are averaged over 30 runs with shading indicating standard error. The learning curve is smoothed over a window of size 10 before averaging across runs.

6 Conclusion

In this work, we proposed a simple approach to learning sparse representations with neural networks, with an activation that naturally produces sparse layers. The activation draws on the idea of binning—or aggregation—which produces a one-hot encoding of an input. This approach ensures sparsity, which can reduce interference when learning online, but suffers from two problems: a loss of precision due to aggregation and zero gradient almost everywhere. We provide a Leaky Tiling Activation (LTA), that ameliorates both of these issues. A parameter η within the LTA provides larger regions with non-negligible gradients, with the positive consequence that it also increases overlap, which can improve discrimination. We show that the LTA still has sparsity guarantees, related to the choice of η. Though sparse representations have been shown to reduce interference when learning online, it has remained difficult to actually learn these sparse representations online. The LTA provides sparse representations by construction, and so it is much easier to use when learning online. We highlight that LTA is robust to high levels of covariate shift, in a synthetic supervised learning problem. We then show across several discrete-action and continuous-action reinforcement learning environments that using the LTA significantly improve learning efficiency and stability, and in most cases even removes the need to use target networks.

REFERENCES

Abadi, M. and et. al. TensorFlow: Large-scale machine learning on heterogeneous systems. *Software available from tensorflow.org*, 2015.

Arpit, D., Zhou, Y., Ngo, H., and Govindaraju, V. Why regularized auto-encoders learn sparse representation? In *International Conference on Machine Learning*, pp. 136–144, 2016.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv*, 2016.

Caselles-Dupré, H., Ortiz, M. G., and Filliat, D. Continual state representation learning for reinforcement learning using generative replay. *arXiv:1810.03880*, 2018.

Chandak, Y., Theocharous, G., Kostas, J., Jordan, S., and Thomas, P. S. Learning Action Representations for Reinforcement Learning. *arXiv:1902.00183*, 2019.

Cover, T. M. Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition. *IEEE Trans. Electronic Computers*, 1965.

French, R. M. Using semi-distributed representations to overcome catastrophic forgetting in connectionist networks. In *Annual Cognitive Science Society Conference*, 1991.

French, R. M. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4): 128–135, April 1999. ISSN 1364-6613. doi: 10.1016/S1364-6613(99)01294-2.

Ghiassian, S., Rafiee, B., Lo, Y. L., and White, A. Improving performance in reinforcement learning by breaking generalization in neural networks. *International Conference on Autonomous Agents and Multi-agent Systems*, 2020.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. *International Conference on Artificial Intelligence and Statistics*, 2010.

Grunwald, G., Hyndman, R., and Tedesco, L. A unified view of linear ar (1) models. 1995.

Heravi, J. R. *Learning Representations in Reinforcement Learning*. PhD thesis, University of California, Merced, 2019.

Hernandez-Garcia, J. F. and Sutton, R. S. Learning sparse representations incrementally in deep reinforcement learning, 2019.

Javed, K. and White, M. Meta-learning representations for continual learning. In *Advances in Neural Information Processing Systems 32*, pp. 1818–1828. Curran Associates, Inc., 2019.

Kemker, R., McClure, M., Abitino, A., Hayes, T. L., and Kanan, C. Measuring catastrophic forgetting in neural networks. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Kim, S., Asadi, K., Littman, M., and Konidaris, G. Deepmellow: Removing the need for a target network in deep q-learning. *International Joint Conference on Artificial Intelligence*, pp. 2733–2739, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

Le, L., Kumaraswamy, R., and White, M. Learning sparse representations in reinforcement learning with sparse coding. In *International Joint Conference on Artificial Intelligence*, pp. 2067–2073, 2017.

LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

Leurent, E. An environment for autonomous driving decision-making, 2018. URL https://github.com/eleurent/highway-env.

Liang, Y., Machado, M. C., Talvitie, E., and Bowling, M. State of the art control of atari games using shallow reinforcement learning. *International Conference on Autonomous Agents & Multiagent Systems*, pp. 485–493, 2016.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2016.

Liu, V., Kumaraswamy, R., Le, L., and White, M. The utility of sparse representations for control in reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 4384–4391, 2019.

Madjiheurem, S. and Toni, L. Representation learning on graphs: A reinforcement learning application. *arXiv:1901.05351*, 2019.

Makhzani, A. and Frey, B. k-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

Makhzani, A. and Frey, B. Winner-take-all autoencoders. In *Advances in Neural Information Processing Systems*, 2015.

McCloskey, M. and Cohen, N. J. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. *Psychology of Learning and Motivation*, 1989.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., and others. Human-level control through deep reinforcement learning. *Nature*, 2015.

Rafati, J. and Noelle, D. C. Learning sparse representations in reinforcement learning. *arXiv:1909.01575*, 2019.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *International Conference on Neural Information Processing Systems*. MIT Press, 1999.

Todorov, E., Erez, T., and Tassa, Y. Mujoco: A physics engine for model-based control. In *International Conference on Intelligent Robots and Systems*, 2012.

van Hasselt, H., Doron, Y., Strub, F., Hessel, M., Sonnerat, N., and Modayil, J. Deep reinforcement learning and the deadly triad. *arXiv:1812.02648*, 2018.

Watkins, C. J. C. H. and Dayan, P. Q-learning. *Machine Learning*, 1992.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. 2017.

Yang, Z., Xie, Y., and Wang, Z. A Theoretical Analysis of Deep Q-Learning. *arXiv:1901.00137*, 2019.

A APPENDIX

This appendix includes the following contents:

1. Section A.1 briefly reviews tile coding which inspires LTA and the naming.
2. Section A.2 shows the proofs for theorems about sparsity guarantee in this paper.
3. Section A.3 discusses possible practical issues when using our LTA function.
4. Section A.4 includes experimental details of Section 5 for reproducible research.
5. Section A.5 presents additional experiments in reinforcement learning setting. It includes: 1) empirical sparsity achieved by LTA; 2) comparison with additional baseline Tile Coding NNs Ghiassian et al. (2020); 3) out of boundary loss in LTA; 4) sensitivity analysis of sparsity control parameter and tile (bin) width; 5) the effect of using LTA in both hidden layers in RL problems.
6. Section A.6 reports the results of LTA on two popular image classification datasets: Mnist LeCun & Cortes (2010) and Mnistfashion Xiao et al. (2017).
7. Section A.7 includes the details of the synthetic supervised learning experiment from Section 4.

A.1 TILE CODING REVIEW

We give a brief review of tile coding here, as tile coding inspires our Leaky Tiling Activation (and its naming).

Tile coding is a generalization of state aggregation, that uses multiple tilings (aggregations) to improve discrimination. For input $z \in [0, 1]$, state-aggregation would map z to a one-hot vector with a 1 in the corresponding bin (bin can be also called tile), with k the number of bins discretized into intervals of length δ. In tile coding, multiple such tilings of the input are concatenated, where each tiling is offset by a small amount. This is depicted in Figure 8, where we show two tilings, one covering from $[-0.05, 1]$ and the other from $[0.0, 1.05]$ both with $k = 4$ and $\delta = 1.05/4 = 0.2625$. The resulting feature vector is a concatenation of these two tilings, to produce 8 features. Two nearby inputs $z = 0.3$ and $z' = 0.55$ would be aggregated together in state-aggregation, making it impossible to produce different values for those inputs. Tile coding, on the other hand, would cause them to share one feature in the second tiling, but each have a distinct feature that enables them to have different values. The large bins/tiles allows for some information sharing, for fast generalization; and the overlapping tilings allows for improved discrimination.

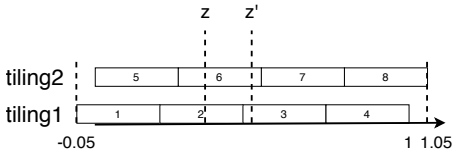


Figure 8: Tile coding maps a scalar to an 8 dimensional binary vector. For example, z activates the second tile on both tilings which gives the feature vector $(0, 1, 0, 0, 0, 1, 0, 0)$. Similarly, $z' \mapsto (0, 0, 1, 0, 0, 1, 0, 0)$.

Unfortunately, tile coding can scale exponentially in the dimension d of the inputs. Each dimension is discretized to a granularity of δ (k regions), with the cross-product between all these regions resulting in k^d bins. One natural approach to make binning more scalable and more practically usable, is to combine it with neural networks (NN). Our activation function—Leaky Tiling Activation—enables a differentiable binning operation.

A.2 PROOFS

We now provide the proof for Proposition 1, Proposition 2, Theorem 1 and Corollary 1 below.

A.2.1 PROOF FOR PROPOSITION 1

For convenience, we define the k-dimensional one-hot vector $\mathbf{e}_i$ whose ith entry is one and otherwise zero.

Proposition 1. *Under Assumption 1, for any $z \in [l, u]$,*

1. *If* $\mathbf{c}_i < z < \mathbf{c}_{i+1}$ *for* $i \in [k-1]$ ***or*** $c_k < z$*, then* $\boldsymbol{\phi}(z) = \mathbf{e}_i$
2. *If* $z = \mathbf{c}_i$ *for* $i \in \{2, ..., k\}$*, then* $\boldsymbol{\phi}(z) = \mathbf{e}_{i-1} + \mathbf{e}_i$
3. *If* $z = \mathbf{c}_1$*, then* $\boldsymbol{\phi}(z) = \mathbf{e}_1$

Proof. In all thee cases, the first max operation in $\boldsymbol{\phi}$ is

$$\max(\mathbf{c} - z, 0) = (0, 0, ..., \mathbf{c}_{i+1} - x, \mathbf{c}_{i+2} - z, ..., \mathbf{c}_k - z).$$

To understand the output of the second max operation, we look at the three cases separately.
Case 1: Because $\mathbf{c}_i < z < \mathbf{c}_{i+1}$, we know that $\mathbf{c}_{i-1} < z - \delta < \mathbf{c}_i$. This implies the second max operation in $\boldsymbol{\phi}$ is:

$$\begin{aligned}&\max(z - \delta - \mathbf{c}, 0)\\ &\quad = (z - \delta - \mathbf{c}_1, z - \delta - \mathbf{c}_2, ..., z - \delta - \mathbf{c}_{i-1}, 0, 0, ..., 0)\end{aligned}$$

Therefore, the sum of both max operations $\max(\mathbf{c} - z, 0) + \max(z - \delta - \mathbf{c}, 0)$ has positive elements everywhere except the ith position, which is zero. Hence $I_+(\max(\mathbf{c} - z, 0) + \max(z - \delta - \mathbf{c}, 0))$ gives a vector where every entry is 1 except the ith entry which is 0. Then $1 - I_+(\max(\mathbf{c} - z, 0) + \max(z - \delta - \mathbf{c}, 0)) = \mathbf{e}_i$.
Case 2: If $z = \mathbf{c}_i, i \in \{2, ..., k\}$, then $\mathbf{c}_{i-1} = z - \delta$, and

$$\begin{aligned}&\max(z - \delta - \mathbf{c}, 0)\\ &\quad = (z - \delta - \mathbf{c}_1, z - \delta - \mathbf{c}_2, ..., z - \delta - \mathbf{c}_{i-2}, 0, ..., 0).\end{aligned}$$

It follows that $\max(\mathbf{c} - z, 0) + \max(z - \delta - \mathbf{c}, 0)$ has exactly two zero entries, at indices $i - 1$ and i. This gives $1 - I_+(\max(\mathbf{c} - z, 0) + \max(z - \delta - \mathbf{c}, 0)) = \mathbf{e}_{i-1} + \mathbf{e}_i$, a vector with ones at indices $i - 1$ and i.
Case 3: When $z = \mathbf{c}_1$, $\max(z - \delta - \mathbf{c}, 0)$ is a zero vector and $\max(\mathbf{c} - z, 0)$ is positive everywhere except the first entry, which is zero. Again this gives $1 - I_+(\max(\mathbf{c} - z, 0) + \max(z - \delta - \mathbf{c}, 0)) = \mathbf{e}_1$. □

A.2.2 Proof for Proposition 2

Proposition 2. *Let* $\mathcal{I}$ *be the set of indices where* $\boldsymbol{\phi}(z)$ *is active (from Theorem 1 we know it can only contain one or two indices). Under Assumption 1, for any* $z \in [l, u]$*, the function* $\boldsymbol{\phi}_\eta(z) = \boldsymbol{\phi}(z) + \Delta$*, where* Δ *is mostly zero, with a few non-zero entries. The vector* Δ *has the following properties:*

1. Δ *is zero at indices* $i \in \mathcal{I}$*, i.e.,* $\boldsymbol{\phi}_\eta(z)$ *equals* $\boldsymbol{\phi}(z)$ *at indices* $i \in \mathcal{I}$*.*
2. Δ *is non-zero at indices* $\{j | j \notin \mathcal{I}, j \in [k], 0 < z - \delta - \mathbf{c}_j \leq \eta, 0 < \mathbf{c}_j - z \leq \eta\}$*.*

Proof. **Part 1.** Let the ith entry be active in $\boldsymbol{\phi}(z)$. We have one of the three cases hold as stated in Theorem 1. Assume $\mathbf{c}_i < z < \mathbf{c}_{i+1}$. Note that

$$\max(\mathbf{c} - z, 0) = (0, 0, ..., \mathbf{c}_{i+1} - z, ..., \mathbf{c}_k - z)$$
$$\max(z - \delta - \mathbf{c}, 0) = (z - \delta - \mathbf{c}_1, ..., z - \delta - \mathbf{c}_{i-1}, 0, ..., 0),$$

taking the sum of the above two equations gives us a vector as following:

$$(z - \delta - \mathbf{c}_1, ..., z - \delta - \mathbf{c}_{i-1}, 0, \mathbf{c}_{i+1} - z, ..., \mathbf{c}_k - z) \tag{6}$$

Then applying $I_{\eta,+}(\cdot)$ to vector equation 6 gives us a vector retaining elements $\leq \eta$ and all other elements become 1. Hence the ith position is zero after applying $I_{\eta,+}(\cdot)$ to vector equation 6. Using one minus this vector would give us a vector with only the ith is one. But this is exactly $\boldsymbol{\phi}_\eta(z)$. And since $\boldsymbol{\phi}_\eta(z) = \boldsymbol{\phi}(z) + \Delta$, and the ith entry of $\boldsymbol{\phi}(z)$ is also one, the ith entry of Δ must be zero. Similar reasoning applies to the cases when $z = \mathbf{c}_i, i \in \{2, ..., k\}$ or $z = \mathbf{c}_1$.

Part 2. Note that applying $I_{\eta,+}(\cdot)$ to the vector $\max(\mathbf{c} - z, 0) + \max(z - \delta - \mathbf{c}, 0)$ keeps all elements no more than η and making all other elements one. As a result, $1 - I_{\eta,+}(\max(\mathbf{c} - z, 0) + \max(z - \delta - \mathbf{c}, 0))$ would give a vector zero everywhere except those entries in $\max(\mathbf{c} - z, 0) + \max(z - \delta - \mathbf{c}, 0)$ which are $\leq \eta$. The set of indices which are $\leq \eta$ in the vector $\max(\mathbf{c} - z, 0) + \max(z - \delta - \mathbf{c}, 0)$ can be written as $\{j | j \in [k], 0 < z - \delta - \mathbf{c}_j \leq \eta, 0 < \mathbf{c}_j - z \leq \eta\}$, which is also the set of indices where $\boldsymbol{\phi}_\eta(z)$ is nonzero. Since $\boldsymbol{\phi}_\eta(z) = \boldsymbol{\phi}(z) + \Delta$ and $\boldsymbol{\phi}(z)$ has nonzero entries in the indices $\mathcal{I}$ and Δ has zero values at those entries by Part 1, then Δ must have nonzero entries at $\{j | j \notin \mathcal{I}, j \in [k], 0 < z - \delta - \mathbf{c}_j \leq \eta, 0 < \mathbf{c}_j - z \leq \eta\}$. □

A.2.3 PROOF FOR THEOREM 1

Theorem 1. Sparsity guarantee for LTA. For any $z \in [l, u], \eta > 0$, $\boldsymbol{\phi}_\eta(z)$ outputs a vector whose number of nonzero entries $\|\boldsymbol{\phi}_\eta(z)\|_0$ satisfies:

$$\|\boldsymbol{\phi}_\eta(z)\|_0 \leq 2\left\lfloor \frac{\eta}{\delta} \right\rfloor + 3$$

Proof. Similar to the proof of Theorem 1, we divide to three cases to prove the result.

Case 1. Consider the case that $\mathbf{c}_i < z < \mathbf{c}_{i+1}, i \in [k-1]$. Note that the number of nonzero entries in $\boldsymbol{\phi}_\eta$ is equal to the number of entries less than η in the vector equation 6, hence we can count the number of entries less than η in $(z - \delta - \mathbf{c}_1, ..., z - \delta - \mathbf{c}_{i-1}, 0, \mathbf{c}_{i+1} - z, \mathbf{c}_{i+2} - z, ..., \mathbf{c}_k - z)$.

First, we count the number of entries that are less than or equal to η on the left side of the ith position. Since the ith position is zero, which indicates $z - \delta - \mathbf{c}_i < 0$, hence $z - \delta - \mathbf{c}_{i-1} - \delta < 0$ and it follows that $0 < z - \delta - \mathbf{c}_{i-1} < \delta$. Then $\delta < z - \delta - \mathbf{c}_{i-1} + \delta = z - \delta - \mathbf{c}_{i-2} < 2\delta$. Hence $(j-1)\delta < z - \delta - \mathbf{c}_{i-j} < j\delta, j \in \{1, ..., i-1\}$. Assume there are m entries $\leq \eta$ on the the left side of the ith. Then $z - \delta - \mathbf{c}_{i-m} \leq \eta$. It follows $(m-1)\delta < \eta$ and hence $m \leq \lfloor \frac{\eta}{\delta} \rfloor + 1$. Hence the total number of elements $\leq \eta$ on the left side of the ith position is at most $\lfloor \frac{\eta}{\delta} \rfloor + 1$.

Second, count the number of entries $\leq \eta$ on the right side of the ith position. Since $\mathbf{c}_i - z < 0$, $0 < \mathbf{c}_i + \delta - z = \mathbf{c}_{i+1} - z < \delta, 0 < \mathbf{c}_{i+2} - z < 2\delta, ...$ Hence the possible number of entries $\leq \eta$ on the right side of ith position is at most $\lfloor \frac{\eta}{\delta} \rfloor + 1$.

As a result, together with the ith position which is $0 \leq \eta$, the possible number of nonzero entries in this case is at most $2\lfloor \frac{\eta}{\delta} \rfloor + 3$.

Case 2. When $z = \mathbf{c}_i, i \in \{2, ..., k\}$, we count the number of entries less than η in the vector

$$\begin{aligned} &\max(\mathbf{c} - z, 0) + \max(z - \delta - \mathbf{c}, 0) \\ &= ((i-1)\delta, ..., 2\delta, \delta, 0, 0, \delta, 2\delta, ..., (k-i)\delta) \end{aligned}$$

Again, we attempt to count the number of entries less than η in this vector by considering $(i-1)\delta, ..., 2\delta, \delta$ and $\delta, 2\delta, ..., (k-i)\delta$ respectively.

We follow the exactly same argument as above and now we have two zero entries at $i-1, i$th positions. The difference is that, the number of entries less than η in the vector $((i-1)\delta, ..., 2\delta, \delta$ can be at most $\lfloor \frac{\eta}{\delta} \rfloor$. As a result, the number of nonzero entries in this case is still at most $2\lfloor \frac{\eta}{\delta} \rfloor + 2$.

Case 3. When $z = \mathbf{c}_1$, $\max(z - \delta - \mathbf{c}, 0)$ is a zero vector and $\max(\mathbf{c} - z, 0) + \max(z - \delta - \mathbf{c}, 0)$ is positive everywhere except the first entry, which is zero. Then we simply count the number of entries $\leq \eta$ on the right side of the 0th position, i.e. $j \in \{1, ..., k\}$. Similar to the analysis in the above Case 1, the possible number of entries $\leq \eta$ on the right side of 0th position is at most $\lfloor \frac{\eta}{\delta} \rfloor$. Hence in this case, the number of nonzero entries is at most $\lfloor \frac{\eta}{\delta} \rfloor + 1$.

In summary, the number of nonzero entries does not exceed $2\lfloor \frac{\eta}{\delta} \rfloor + 3$. This completes the proof. □

Remark. As we empirically demonstrated in Table 1 and below Figure 10 and Figure 9, the actual sparsity achieved by LTA is lower than the upper bound. This is because our upper bound is for any possible input z. Consider that in Case 1 in the above proof, we count the number of entries that are less than or equal to η on the left side and right side of the ith position. There are $\lfloor \frac{\eta}{\delta} \rfloor + 1$ such entries on both sides only when z is exactly equal to $\frac{\mathbf{c}_i + \mathbf{c}_{i-1}}{2}$, which is unlikely to happen in practice.

A.2.4 PROOF FOR COROLLARY 1

Corollary 1 *Let $\rho \in [0, 1)$ be the desired sparsity level: the maximum proportion of nonzero entries of $\boldsymbol{\phi}_\eta(z), \forall z \in [l, u]$. Assume $\rho k \geq 3$, i.e., some inputs have three active indices or more (even with $\eta = 0$, this minimal active number is 2). Then η should be chosen such that*

$$\left\lfloor \frac{\eta}{\delta} \right\rfloor \leq \frac{k\rho - 3}{2} \quad \textit{or equivalently} \quad \eta \leq \frac{\delta}{2}\left(\lfloor k\rho \rfloor - 1\right) \tag{7}$$

Proof. Because $\|\boldsymbol{\phi}_\eta(z)\|_0 \leq \lfloor k\rho \rfloor$, from Theorem 1 it is sufficient to pick η such that $2\lfloor \frac{\eta}{\delta} \rfloor + 3 \leq \lfloor k\rho \rfloor \leq k\rho$. This gives $\lfloor \frac{\eta}{\delta} \rfloor \leq (\lfloor k\rho \rfloor - 3)/2 \leq (k\rho - 3)/2$. Additionally, we know $\frac{\eta}{\delta} - 1 \leq \lfloor \frac{\eta}{\delta} \rfloor \leq (\lfloor k\rho \rfloor - 3)/2$, giving the second inequality. □

A.3 More discussion about LTA

Our development of the LTA assumed that the inputs are bounded in the range $[l, u]$ (recall that $\mathbf{c} = [l, l+\delta, l+2\delta, ..., u]$). This is not guaranteed for the standard inputs to activations, namely $z = \mathbf{x}^\top \mathbf{w}$ for some weight vector $\mathbf{w}$. The LTA can still be used if $z \notin [l, u]$, but then the gradient of the LTA will be zero. This means that the weights $\mathbf{w}$ cannot be adjusted for inputs where z becomes too large or too small. This issue is usually called *gradient vanish*, which is common in many popular, existing activation functions such as ReLU, tanh, sigmoid, etc.

Our LTA has two simple ways to avoid this issue: 1) use a squashing activation before handing z to the LTA and 2) regularize (penalize) z that falls outside the range $[l, u]$. For example, tanh can be applied first to produce $z = \tanh(\mathbf{x}^\top \mathbf{w}) \in [-1, 1]$. Though the simplest strategy, using tanh function can be problematic in deep neural networks due to gradient vanishing problems. An alternative is to use a penalty, that pushes out-of-boundary z back into the chosen range $[-u, u]$

$$r(z) \stackrel{\text{def}}{=} I(|z| > u) \circ |z| \tag{8}$$

This penalty is easily added to the loss, giving gradient $\frac{\partial r(z)}{\partial z} = I(z > u) - I(z < -u)$. For example, $z = \max(\mathbf{x}^\top \mathbf{w}, 0)$, which might produce a value greater than u. If $z > u$, then the gradient pushes the weights $\mathbf{w}$ to decrease z. It should be noted that the number of nonzero entries can only decrease when z goes out of boundary. However, in our experiments in our main paper, we found such out of boundary loss is unnecessary on all of our tested domains. Furthermore, we further verify that our LTA performs stably across different weights for the above regularization in Section A.5.

A.4 Reproducible Research

Common settings. All discrete action domains are from OpenAI Gym Brockman et al. (2016) with version 0.14.0. Deep learning implementation is based on tensorflow with version 1.13.0 Abadi & et. al (2015). We use Adam optimizer Kingma & Ba (2015), Xavier initializer Glorot & Bengio (2010), mini-batch size $b = 64$, buffer size 100k, and discount rate $\gamma = 0.99$ across all experiments. We evaluate each algorithm every 1k training/environment time steps.

Algorithmic details. We use 64×64 ReLU units neural network for DQN and 200×100 ReLU units neural network for DDPG. All activation functions are ReLU except: the output layer of the Q value is linear. The weights in output layers were initialized from a uniform distribution $[-0.003, 0.003]$. Note that we keep the same LTA setting across *all* experiments: we set $[l, u] = [-20, 20]$; we set $\delta = \eta = 2.0$, $\mathbf{c} = \{-20, -18, -16, ..., 18\}$, and hence $k = 40/2 = 20$. This indicates that the DQN-Large and DDPG-Large versions have $64 \times 20 = 1280$ and $100 \times 20 = 2000$ ReLU units in the second hidden layer. For RBF coding, we set the bandwidth as $\sigma = 2.0$, and uses the same tiling (i.e. $\mathbf{c}$ vector) as our LTA.

Meta-parameter details. For DQN, the learning rate is 0.0001 and the target network is updated every 1k steps. For DDPG, the target network moving rate is 0.001 and the actor network learning rate is 0.0001, critic network learning rate is 0.001. For l_1, l_2 regularization variants, we optimize its regularization weight from $\{0.1, 0.01, 0.001, 0.0001\}$ on MountainCar, then we fix the chosen optimal weight 0.01 across all domains.

Environmental details on discrete control domains. We set the episode length limit as 2000 for MountainCar and keep all other episode limit as default settings. We use warm-up steps 5000 for populating the experience replay buffer before training. Exploration noise is 0.1 without decaying. During policy evaluation, we keep a small noise $\epsilon = 0.05$ when taking action.

Environmental details on continuous control domains. On Mujoco domains, we use default settings for maximum episodic length. We use $10,000$ warm-up time steps to populate the experience replay buffer. The exploration noise is as suggested in the original paper by Lillicrap et al. (2016).

A.5 ADDITIONAL EXPERIMENTS ON REINFORCEMENT LEARNING PROBLEMS

This section shows the following additional empirical results which cannot be put in the main body due to space limitations.

- A.5.1 empirically investigates the representation sparsity generated by LTA.
- A.5.2 shows the comparison with Tile Coding NNs (Ghiassian et al., 2020), which first tile code inputs before feeding them into a neural network.
- A.5.3 shows the empirical results of DQN-LTA using linear activation function with regularization weights $\in \{0.0, 0.01, 1.0\}$ for the out of bound loss in Eq 8.
- A.5.4 shows the sensitivity of LTA to the sparsity control parameter η and tile width parameter δ.
- A.5.5 shows the results when applying LTA to both hidden layers rather than just the second hidden layer in RL problems.

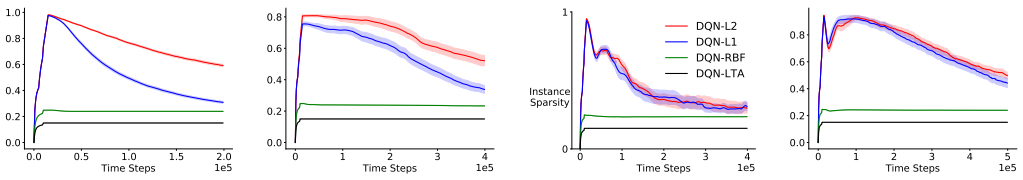
A.5.1 EMPIRICALLY MEASURING SPARSITY

We also report two sparsity measures of the learned representations averaged across all steps in Table 1. We compute an estimate of sparsity by sampling a mini-batch of samples from experience replay buffer and taking the average of them. Instance sparsity corresponds to proportion of nonzero entries in the feature vector for each instance. Overlap sparsity French (1991) is defined as $overlap(\phi, \phi') = \sum_i I(\phi_i \neq 0) I(\phi'_i \neq 0)/(kd)$, given two kd-dimensional sparse vectors ϕ, ϕ'. Low overlap sparsity potentially indicates less feature interference between different input samples. Theorem 1 guarantees that the instance sparsity, with LTA, should be no more than 12.5% for our setting of $k = 40, \eta = \delta$; and should be no more than 25% when $k = 20, \eta = \delta$. In the table, we can see that LTA has lower (better) instance sparsity than the upper bound provided by the theorem. Further, LTA achieves the lowest instance and overlap sparsity among all baselines in an average sense.

Figure 9 and Figure 10 are corresponding to the learning curves as shown in Figure 6 in Section 5.3. We show learning curves of instance/overlap sparsity as a function of training steps for LTA, l_1, l_2 regularization and RBF. The instance sparsity is computed by randomly sampling a mini-batch of states from replay buffer and count the average number of nonzero entries in that mini-batch divided by feature dimension kd. The overlap sparsity is computed by randomly sampling two mini-batches of states from the replay buffer and count the average number of simultaneously activated (i.e. nonzero) entries in both mini-batches. From the sparsity learning curves, one can see that our LTA has very stable sparsity since the beginning of the learning, and the sparsity is almost constant cross domains. Particularly, the overlap sparsity is very low, which possibly indicates low representation interference.

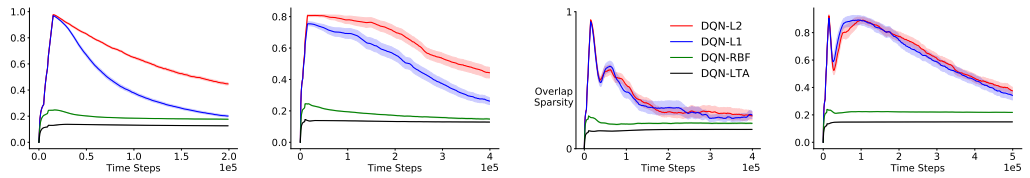
Table 1: Sparsity on LunarLander (average across time steps)

Sparsity	LTA(k=40)	LTA(k=20)	L1	L2	RBF
Instance	7%	14%	16%	44%	99%
Overlap	4%	8%	10%	34%	99%



(a) Acrobot-v1 (b) MountainCar-v0 (c) CartPole-v1 (d) LunarLander-v2

Figure 9: Instance sparsity v.s. number of time steps on MountainCar, CartPole, Acrobot, LunarLander. The results are averaged over 20 random seeds and the shade indicates standard error.



(a) Acrobot-v1 (b) MountainCar-v0 (c) CartPole-v1 (d) LunarLander-v2

Figure 10: Overlap sparsity (number of simultaneously activated entries in the sparse feature vectors) v.s. number of time steps by averaging over 20 random seeds and the shade indicates standard error.

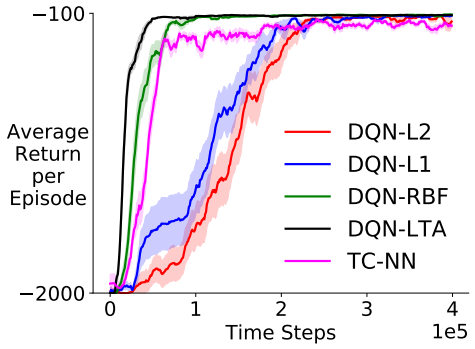


Figure 11: DQN-LTA compares with TCNN. Evaluation learning curves are averaged over 20 random seeds and the shade indicates standard error. All variants are trained without target networks.

A.5.2 Comparison with TC-NN

We include the figure of comparing with TC-NN from the work by Ghiassian et al. (2020), which use a regular tile coding to process the input before feeding into the neural network. It should be noted that this method itself is not a sparse representation learning technique, and requires the original observation space to be scaled within $[0, 1]$. Figure 11 shows that TC-NN does indeed improve performance for DQN, but not compared to LTA. DQN-LTA both learns faster and reaches a better, more stable solution. In our TC-NN implementation, as a correspondent to LTA, we use binning operation to turn each raw input variable to 20 binary variables through binning operation. We use the same learning rate 0.0001 as other baselines.

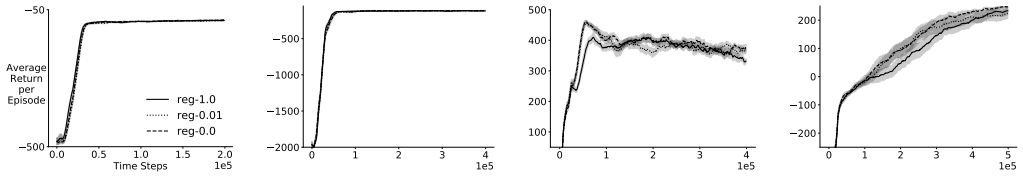
A.5.3 Insensitivity to out of boundary loss

We demonstrate the effect of using an out of boundary loss as discussed in Section A.3. This is supplementary to our experiments in the main body, where we did not use an out of boundary loss, i.e. the regularization weight is 0. To highlight the issue of out of boundary, we intentionally use a tiling vector $\mathbf{c}$ with a small range $[-1, 1]$ with 20 tiles, i.e. $\delta = 2/20 = 0.1$ and set the sparsity control parameter $\eta = \delta = 0.1$. In Figure 12, one can see that our algorithm typically does not need such a regularization even if we use a small tiling.

A.5.4 Sensitivity to Tile Width and Sparsity

We empirically investigate the sensitivity of LTA to its two parameters: tile width δ and sparsity control parameter η. We use LunarLander, because it was the most difficult discrete-action environments tested. For both η and δ, we sweep over $0.8/2^i, i \in \{0, 1, 2, ..., 8\}$. Note that this range is extreme ($\delta = 0.8$ gives two bins, and $\eta = 0.8$ significantly increases overlap); we do so to see a clear pattern.

Figure 13 shows the early learning performance of all possible $9 \times 9 = 81$ combinations of η, δ. We report the average episodic return (rounded to integers) within the first 300k time steps, for each combination. The results are averaged over 5 runs. The pattern is as expected. 1) The algorithms



(a) Acrobot-v1 (b) MountainCar-v0 (c) CartPole-v1 (d) LunarLander-v2

Figure 12: DQN-LTA trained with linear activation function with different regularization weight (-reg-1.0 means regularization is set as 1.0). One can see that our LTA is *not sensitive to this choice* as those learning curves are almost overlapping with each other. Evaluation learning curves are averaged over 10 random seeds and the shade indicates standard error. All variants are trained without target networks.

performs best with a reasonably small δ and η. 2) For extremely small η and δ, performance is poor, as expected because the gradient is very small and so learning is slow. 3) Given a fixed tile width δ, the performance degrades as η becomes smaller than δ, since again the activation has many regions with zero derivative. 4) If η gets too large, the sparsity level increases and again performance degrades, though performance remained quite good for a broad range between 0.025 and 0.2. 5) If δ gets large (big bins), performance degrades more so that with larger η, which matches the intuition that η provides overlap rather than just increasing bin size and so losing precision.

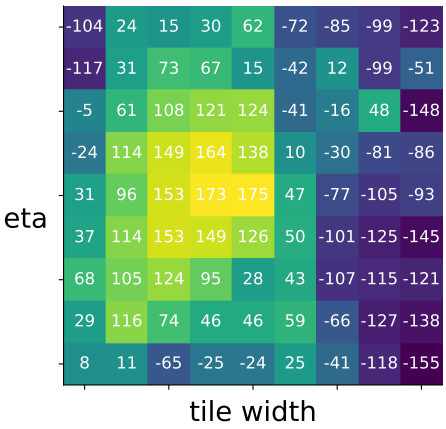
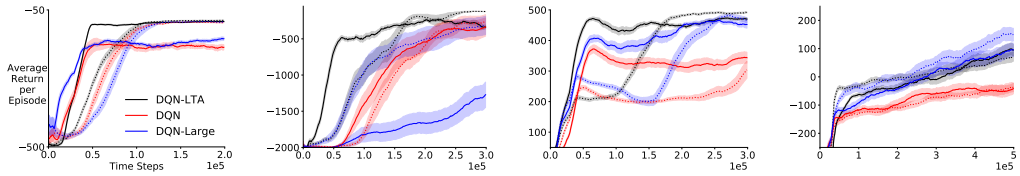


Figure 13: Sensitivity of η, δ on LunarLander-v2.

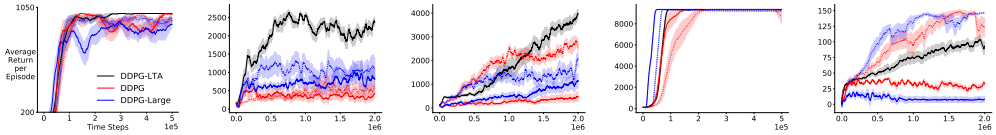
A.5.5 Empirical results when using LTA to both hidden layers

As an activation function, LTA can be naturally applied in any hidden layer in a NN. As a convention, we typically use the same activation functions in a fully connected NN. In this section, we present such results on the reinforcement learning domains, as a supplement to the results in Section 5. All setting is the same as those in Section 5, except that we apply LTA to both hidden layers in DQN-LTA and DDPG-LTA. It should be noted that, in this case, DQN-LTA and DDPG-LTA have the same number of training parameters as their -Large correspondents respectively. Figure 14 shows the results on the discrete domains, and Figure 15 shows the results on the continuous control domains. It can be seen that, the LTA versions are better than ReLu versions in the case of not using a target network. This further validates the utility of our LTA in dealing with nonstationary problems.



(a) Acrobot (v1) (b) MountainCar (v0) (c) CartPole (v1) (d) LunarLander (v2)

Figure 14: Evaluation learning curves showing episodic return versus environment time steps. The **dotted** line indicates algorithms trained *with* target networks. The results are averaged over 20 runs and the shading indicates standard error. The learning curve is smoothed over a window of size 10 before averaging across runs.



(a) InvertedPendu (b) Hopper (c) Walker2d (d) InvertedDouble (e) Swimmer

Figure 15: Evaluation learning curves on Mujoco environments, averaged over 20 runs with shading indicating standard error. All algorithms are trained without target networks. The learning curve is smoothed over a window of size 30 before averaging across runs.

A.6 Results on Image Classification Tasks

We now report the empirical results on two popular image classification tasks: Mnist (LeCun & Cortes, 2010) and Mnistfashion (Xiao et al., 2017). We found that LTA does not present any clear advantage or disadvantage in such a conventional supervised learning setting.

On Mnist, LTA achieves testing error 1.22%, and ReLu achieves 1.38%. On Mnistfashion, LTA achieves testing error 10.67%, and ReLu achieves 10.87%.

Details. The NN architecture we use is two convolution layer followed by two fully connected 32×32 layers. The first convolutional layer has 6 filters with size 5×5 and is followed by a max pooling operation. The second convolutional layer has 16 filters with the same size followed by a max pooling. LTA is applied to the second hidden layer and the setting is exactly the same as we used in the RL experiment A.4. To optimize learning rate, we optimize over the range $\{0.00003, 0.00001, 0.0003, 0.0001, 0.003, 0.001\}$. We use 10-fold cross validation to choose the best learning rate and the above error rate is reported on testing set by using the optimal learning rate at the end of learning. The standard deviation of the testing error (randomness comes from NN initialization and random shuffling) is sufficiently small to get ignored.

A.7 Construction of Piecewise Random-Walk Problem

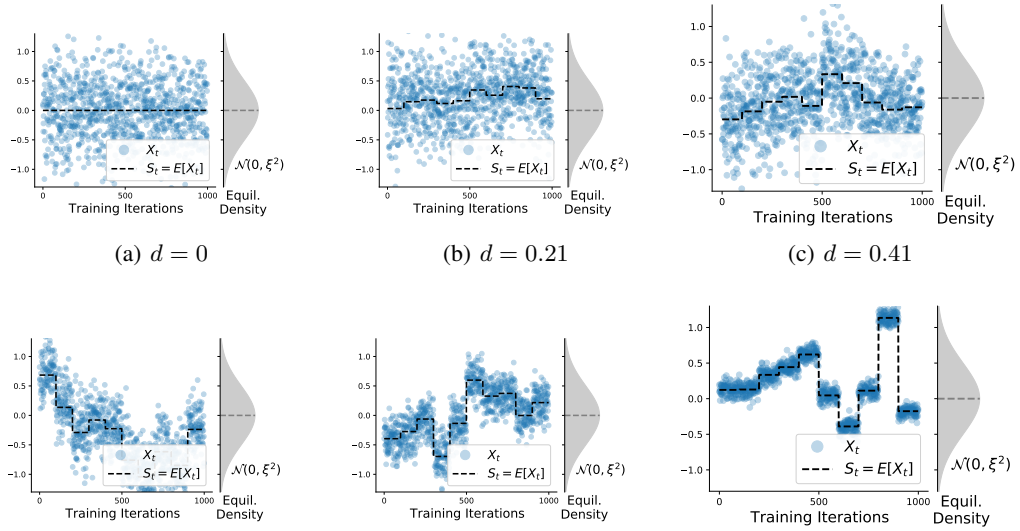
In Section 4, recall that we train with data generating process $\{(X_t, Y_t)\}_{t\in\mathbb{N}}$ so that interference difficulty is controllable, but permitting fair comparison via a fixed equilibrium distribution.

The temporal correlation in $\{X_t\}_{t\in\mathbb{N}}$ is designed to correspond to state space trajectories through MDPs with a high degree of local state space connectedness. e.g. an agent can only move to adjacent cells in GridWorld, the paddle and ball in Atari Breakout can only move so far in a single frame, and most successor states in Go only differ by a few pieces. Figure 16 depicts sample trajectories across a range of difficulties, alongside the fixed equilibrium distribution.

Here we rigorously construct $\{X_t\}_{t\in\mathbb{N}}$ and $\{S_t\}_{t\in\mathbb{N}}$, and show they have the claimed properties.

Let random variable $E_t \sim \mathcal{N}(\epsilon; 0, \sigma^2)$ be a source of Gaussian noise for the recursive random variable S_t

$$S_{t+1} = (1-c)S_t + E_t \tag{9}$$



(a) $d = 0$ (b) $d = 0.21$ (c) $d = 0.41$

(d) $d = 0.62$ (e) $d = 0.83$ (f) $d = 0.98$

Figure 16: Sample trajectories of $\{X_t\}_{t\in\mathbb{N}}$ (blue) and $\{S_t\}_{t\in\mathbb{N}}$ (black) across a range of difficulties. Note in particular the fixed equilibrium distribution (gray) and iid sampling for $d = 0$ (top left).

If $c \in (0, 1]$, then $\{S_t\}_{t\in\mathbb{N}}$ is a linear Gaussian first order autoregressive process, denoted AR(1).[2] Then the *equilibrium distribution* of $\{S_t\}_{t\in\mathbb{N}}$ is also Gaussian. If S_0 is sampled from the equilibrium distribution, then the process $\{S_t\}_{t\in\mathbb{N}}$ is stationary with $E[S_t] = 0$ and variance

$$\nu^2 \doteq E[S_t^2] = \frac{\sigma^2}{2c - c^2} \tag{10}$$

It also follows that $\{S_t\}_{t\in\mathbb{N}}$ is ergodic. See Grunwald et al. (1995) for reference on AR(1) processes, especially the discussion of equation (3.3) for linear Gaussian AR(1) processes, and section 5.3 for their stationarity and ergodicity.

Let s_t be a realization of S_t. We define $X_t|S_t = s_t$ as a Gaussian r.v. with mean s_t and variance β^2:

$$X_t|S_t = s_t \sim \mathcal{N}(x; s_t, \beta^2) \tag{11}$$

The process $\{X_t\}_{t\in\mathbb{N}}$ is also a linear Gaussian AR(1) process, so we can again rely on established AR(1) process theory to conclude that the *equilibrium distribution* of $\{X_t\}_{t\in\mathbb{N}}$ is Gaussian. Moreover, if X_0 is sampled from the equilibrium distribution, then the process $\{X_t\}_{t\in\mathbb{N}}$ is stationary with $E[X_t] = 0$ and variance ξ^2

$$\xi^2 \doteq E[X_t^2] = \beta^2 + \nu^2 = \beta^2 + \frac{\sigma^2}{2c - c^2} \tag{12}$$

Refer to Theorem 4 in Section A.7.2 for proof that $\{X_t\}_{t\in\mathbb{N}}$ is a linear Gaussian AR(1) process, and that the resulting equilibrium distribution has variance specified by equation 12. Note there are several variance terms in the above construction.

- σ^2 is the noise variance in the random walk, i.e. the source of jumps in $\{S_t\}_{t\in\mathbb{N}}$
- ν^2 is the equilibrium distributon variance of $\{S_t\}_{t\in\mathbb{N}}$
- β^2 is the variance in any particular X_t given s_t at time step t

[2] If $c = 0$, then S_t is a simple Gaussian random walk. We do not use such a process, because sample paths are unbounded, hitting $\pm\sqrt{t}$ infinitely often as $t \to \infty$

- ξ^2 is the equilibrium distribution variance of $\{X_t\}_{t\in\mathbb{N}}$

The processes $\{S_t\}_{t\in\mathbb{N}}$ and $\{X_t\}_{t\in\mathbb{N}}$ are fully determined by the quantities c, σ^2, β^2. For our experiments we wish to have a single parameter d such that the equilibrium distribution of $\{X_t\}_{t\in\mathbb{N}}$ is fixed for all $d \in [0, 1)$, but temporal correlation is variable. Also, $d = 0$ should correspond to zero temporal correlation. The following two theorems specify this behaviour more rigorously.

Theorem 2. *Let difficulty parameter $d \in [0, 1)$ and $\xi^2 = (\frac{B}{2})^2$ for some $B \in \mathbb{R}^+$. Then $\{X_t\}_{t\in\mathbb{N}}$ will have fixed equilibrium distribution $\mathcal{N}(0, \xi^2)$, invariant to d, if parameters c, σ^2, β^2 are set as follows*

$$c = 1 - \sqrt{1-d}$$
$$\sigma^2 = d^2(\frac{B}{2})^2$$
$$\beta^2 = (1-d)(\frac{B}{2})^2$$

ξ^2 is defined in terms of $B \in \mathbb{R}^+$ simply because B is a more intuitive design parameter. In particular, B is a high probability bound with $P(X_t \in [-B, B]) \leq 0.95$ for all t, since $\{X_t\}_{t\in\mathbb{N}}$ is ergodic.

Theorem 3. *$d = 0$ and $S_0 = 0$ induces iid X_t from $\mathcal{N}(x; 0, \xi^2)$, the equilibrium distribution of $\{X_t\}_{t\in\mathbb{N}}$.*

We defer proof of both theorems to Section A.7.2.

Correlation difficulty d is intuitively characterized as follows.

- As d increases, β^2 decreases, so a smaller portion of the overall state space is supported at any given time. i.e. $P(X_t = x|S_t = s_t)$ becomes increasingly narrow, inducing higher temporal correlation.
- As d increases, the noise in random walk $\{S_t\}_{t\in\mathbb{N}}$ increases in amplitude, so that larger jumps in the mean of any particular X_t are more likely.
- At $d = 0$, all correlation difficulty from $\{X_t\}_{t\in\mathbb{N}}$ has been removed, in that we recover iid sampling.
- As $d \to 1$, $\{X_t\}_{t\in\mathbb{N}}$ converges towards pathological correlation, with $X_t \xrightarrow{d} \delta(s_t)$, where $\delta(\cdot)$ is the Dirac delta distribution. i.e. X_t becomes constant everywhere except the jumps in realization s_t of S_t.

Despite the above dynamics, the equilibrium distribution $\mathcal{N}(x; 0, \xi^2)$ is identical for all difficulty settings $d \in [0, 1)$, because the increase (or decrease) in parameters σ^2, c is tuned specifically to counteract the decrease (or increase) in β^2. Having identical equilibrium distributions means that a hypothetical ideal algorithm, capable of perfectly mitigating interference, would train the identical approximator f_θ for any correlation difficulty $d \in [0, 1)$. Hence, we use approximation loss on the equilibrium distribution as a measure of robustness to interference.

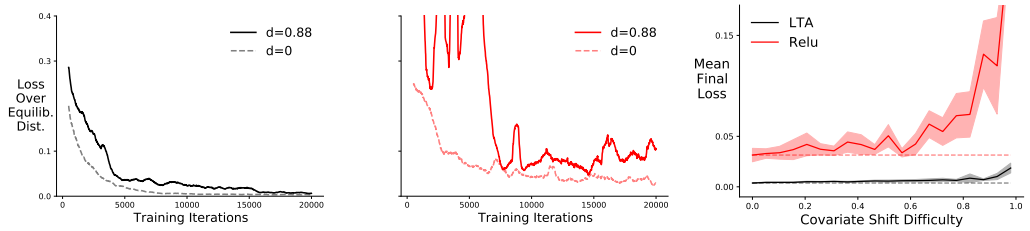
To summarize, correlation difficulty $d \in [0, 1)$ controls the likelihood of interference-inducing samples throughout training, but also permits fair comparison between values of d by the (fixed) equilibrium distribution of $\{X_t\}_{t\in\mathbb{N}}$.

A.7.1 Hyperparameter Selection for Controlled Interference Experiments

Figure 17 compares the performance of LTA and ReLU across a range of difficulty settings, where each of the LTA and ReLU networks are of two hidden layers. First, network were sizes chosen (with hidden layers of equal sizes) such that loss on the iid setting ($d = 0$) was minimized. Next, learning rate is optimized for each difficulty setting, so that the best performance on all networks is reported. Figure 17(c) was included in the main body; we include it again here, to show it with the learning curves. Figure 18 shows learning rate sensitivity curves for the LTA and ReLU networks for three different difficulty settings. The ADAM optimizer was used all experiments.

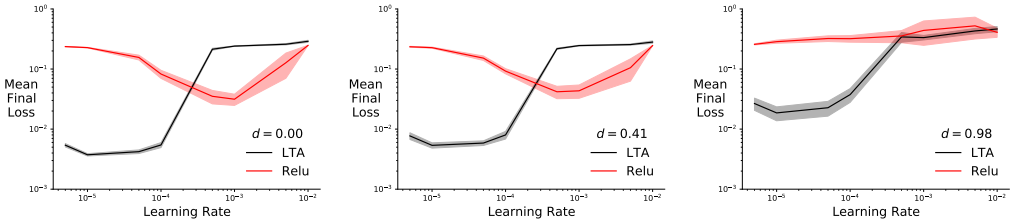
When searching network sizes such that performance on the iid setting was optimized, we found that the ReLU network required fewer parameters than the comparable LTA network. (67K vs 5.2K

learnable parameters.) In order to verify that the difference in parameter count was not a significant factor in our experiments, we repeated all runs with a significantly overparametrized network (same architecture, but wider hidden layers) of 81K learnable parameters. As indicated by Figure 19, the larger ReLU network performs very similarly to the smaller one—in most cases agreeing within $p = 0.05$.



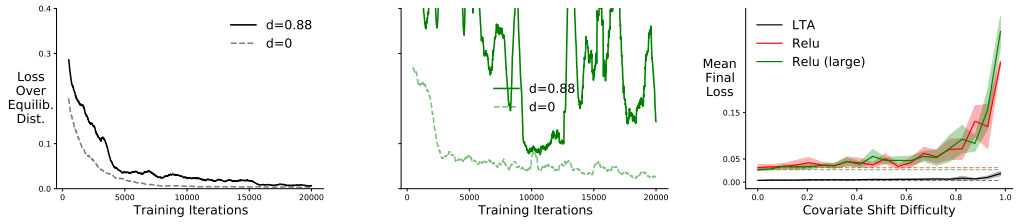
(a) LTA (b) ReLU (c) ReLU and LTA, $d \in [0, 1)$

Figure 17: Left, Middle: Learning curve of loss over stationary distribution during training on low difficulty (dotted) and high difficulty (solid) settings for a two layer neural nets. The curves are smoothed over a window of 50. Right: The final loss over stationary distribution after 20K training iterations across a range of difficulty settings, shown as the mean of 10 runs with the shaded region corresponding $p = 0.05$. The final loss per run is computed as the mean over the final 2.5K iterations, with the iid setting $d = 0$ (dotted) shown for baseline comparison.



(a) Lowest Difficulty (b) Moderate Difficulty (c) High Difficulty

Figure 18: Learning rate sensitivity of LTA and ReLU for iid, mildly correlated, and severely correlated X_t (left, middle, right, respectively.) These curves corroborate our findings that, in general, LTA prefers lower learning rates (but converges more quickly nonetheless.)



(a) LTA (b) ReLU-large (c) ReLU-large and LTA, $d \in [0, 1)$

Figure 19: Left, Middle: Learning curve of loss over stationary distribution during training on low difficulty (dotted) and high difficulty (solid) settings for two layer neural nets, both having similar numbers of learnable parameters. The curves are smoothed over a window of 50. Right: The final loss over stationary distribution after 20K training iterations across a range of difficulty settings, shown as the mean of 10 runs with the shaded region corresponding $p = 0.05$. The final loss per run is computed as the mean over the final 2.5K iterations, with the iid setting $d = 0$ (dotted) shown for baseline comparison.

A.7.2 PROOFS FOR COVARIATE SHIFT PROPERTIES

Theorem 4. *Let $\{E_t\}_{t\in\mathbb{N}}, \{S_t\}_{t\in\mathbb{N}}, \{X_t\}_{t\in\mathbb{N}}$ be stochastic processes such that*

$$E_t \sim \mathcal{N}(\epsilon; 0, \sigma^2)$$
$$S_{t+1} = (1-c)S_t + E_t \tag{13}$$
$$X_t | S_t = s_t \sim \mathcal{N}(x; s_t, \beta^2) \tag{14}$$

where $c \in (0, 1]$ and s_t is a realization of S_t. Then $\{X_t\}_{t\in\mathbb{N}}$ is a linear Gaussian AR(1) process with equilibrium distribution $\mathcal{N}(x; 0, \beta^2 + \frac{\sigma^2}{2c-c^2})$.

Proof. Begin by rewriting equation 14 to give X_t as a sum of S_t and a mean zero Gaussian r.v. B_t

$$X_t = S_t + B_t \qquad\qquad B_t \sim \mathcal{N}(0, \beta^2), \text{ iid} \tag{15}$$

Now substitute according to 13

$$X_t = (1-c)S_{t-1} + E_{t-1} + B_t \tag{16}$$

Let Gaussian r.v. $\Theta \sim \mathcal{N}(0, \alpha^2)$ with some variance α^2.

E_{t-1}, B_t are independent Gaussian, so $E_{t-1} + B_t$ can be written as a linear combination of B_{t-1}, Θ_t, since all are independent. Specifically, fix α^2 so that the following holds

$$E_{t-1} + B_t = (1-c)B_{t-1} + \Theta_t \tag{17}$$

and substitute into equation 16

$$\begin{aligned} X_t &= (1-c)S_{t-1} + (1-c)B_{t-1} + \Theta_t \\ &= (1-c)(S_{t-1} + B_{t-1}) + \Theta_t \\ &= (1-c)X_{t-1} + \Theta_t \end{aligned}$$

By inspection, $\{X_t\}_{t\in\mathbb{N}}$ is a linear Gaussian AR(1) process with coefficient $(1-c)$ and noise variance α^2. Elementary AR process theory gives the equilibrium distribution as

$$\mathcal{N}\left(0, \frac{\alpha^2}{1-(1-c)^2}\right) \tag{18}$$

For 17 to hold, we need the first and second moments of LHS and RHS to be equal. All terms are mean zero Gaussian, so it suffices to show when the LHS and RHS have equal variance:

$$\begin{aligned} \mathrm{Var}(E_{t-1} + B_t) &= \mathrm{Var}((1-c)B_{t-1} + \Theta_t) \\ \sigma^2 + \beta^2 &= (1-c)^2\beta^2 + \alpha^2 \\ \alpha^2 &= \sigma^2 + \beta^2 - (1-c)^2\beta^2 \\ \alpha^2 &= \sigma^2 + (1-(1-c)^2)\beta^2 \end{aligned}$$

Substituting into 18

$$
\begin{aligned}
&\mathcal{N}\left(0, \frac{\alpha^2}{1-(1-c)^2}\right) \\
=&\mathcal{N}\left(0, \frac{\sigma^2+(1-(1-c)^2)\beta^2}{1-(1-c)^2}\right) \\
=&\mathcal{N}\left(0, \frac{\sigma^2}{1-(1-c)^2}+\beta^2\right) \\
=&\mathcal{N}\left(0, \frac{\sigma^2}{2c-c^2}+\beta^2\right)
\end{aligned}
$$

So $\{X_t\}_{t\in\mathbb{N}}$ is linear Gaussian AR(1) with the desired equilibrium distribution.

□

Theorem 2

Let difficulty parameter $d \in [0, 1)$ and $\xi^2 = (\frac{B}{2})^2$ for some $B \in \mathbb{R}^+$. Then $\{X_t\}_{t\in\mathbb{N}}$ will have fixed equilibrium distribution $\mathcal{N}(0, \xi^2)$, invariant to d, if parameters c, σ^2, β^2 are set as follows

$$
\begin{aligned}
c &= 1-\sqrt{1-d} \\
\sigma^2 &= d^2(\frac{B}{2})^2 \\
\beta^2 &= (1-d)(\frac{B}{2})^2
\end{aligned}
$$

Proof. By the above theorem 4, $\xi^2 = \beta^2 + \frac{\sigma^2}{2c-c^2} = (1-d)(\frac{B}{2})^2 + d^2(\frac{B}{2})^2/d = (\frac{B}{2})^2$. This completes the proof. □

Theorem 3

$d = 0$ and $S_0 = \delta(0)$ induces iid X_t from $\mathcal{N}(x; 0, \xi^2)$, the equilibrium distribution of $\{X_t\}_{t\in\mathbb{N}}$.

Proof.

$$
\begin{aligned}
S_{t+1} &= (1-c)S_t + E_t && (\textit{by } 9) \\
&= S_t + E_t && (d=0 \implies c=0) \\
&= S_t && (d=0 \implies \sigma^2=0 \implies E_t \sim \delta(0))
\end{aligned}
$$

Hence, $d = 0$ implies S_t is constant over all time t, and $S_0 = \delta(0)$ gives $S_t = \delta(0)$ for all time t.

Let r.v. $B_t \sim \mathcal{N}(0, \beta^2)$ i.i.d., then $X_t = S_t + B_t = B_t$ when S_t is Dirac delta concentrating at zero. $\beta^2 = (1-d)(\frac{B}{2})^2 = (\frac{B}{2})^2$ by the setting from Theorem 2 and setting $d = 0$. Hence, $d = 0$ induces iid sampling from the equilibrium distribution of X_t. This completes the proof. □